Pseudo-likelihood Inference for Gaussian Markov Random Fields

Tom Burr*1, Alexei Skurikhin2

Statistical Sciences/Los Alamos National Laboratory USA, Space Data Systems/Los Alamos National Laboratory USA

*1tburr@lanl.gov; 2alexei@lanl.gov

Abstract

Gaussian Markov random fields (GMRFs) are an important example of MRFs with many applications, particularly because GMRFs are known to provide effective approximations to any MRF. Despite their relative computational simplicity, inference in GMRFs using maximum likelihood (ML) is not always feasible. Therefore, this paper compares the inference quality using the pseudolikelihood, a well-known computational shortcut to full ML, and in addition the generalized lambda distribution is simulated to investigate robustness to departure from the Gaussian distribution.

Keywords

Gaussian Markov Random Fields; Pseudolikelihood; Robustness

Introduction

Spatial data are often collected over a grid such as an agricultural plot, a census tract, or at each pixel in an image. For example, Figure 1 is well known wheat yield data (Gelfand et al., 2010) first reported in 1911 from a yield uniformity trial at Rothamsted agricultural station. The celebrated Hammersley-Clifford theorem (Gelfand et al., 2010) specifies the most general possible joint probability distribution of such spatial data indexed by the grid points.

Markov random fields (MRFs) are a useful type of joint probability function that obey a conditional independence property and thus can make inference more tractable. MRFs model undirected probabilistic interactions among the variables. Unlike joint probability distributions that simultaneously specify distributions across the entire field, MRFs are based on a collection of conditional distributions that rely on local neighborhoods of each element. This local dependence provides several advantages, including computational tractability (relative to more general random fields) and model extensions to account for non-stationarity, discontinuity, and varying spatial properties at various scales of resolution which are easily accessible in the MRF framework.

Gaussian Markov random fields (GMRFs) are an important example of MRFs with many applications, particularly because GMRFs are known to provide effective approximations to any MRF (Rue and Held, 2005). The estimated marginal probability density of the 500 wheat yields (ranging from 2.73 to 5.16 pounds, in 20 rows and 25 columns covering 1 a acre uniform agricultural plot) from Figure 1 and shown to be Gaussian-like in Figure 2 was estimated using kernel density estimation (Hastie et al., 2001) implemented in R (R, 2004).

GMRFs have been applied to spatial data including maps of disease incidence and rainfall, to image data, and to many other data types that are collected over a grid (Rue and Held, 2005). Although vector notation may be utilized for simplicity, the spatial index is often two-dimensional, and the vector x is a vector representation of data on a grid such as horizontal and vertical pixels in an image.

Despite their relative computational simplicity, inference in GMRFs using maximum likelihood (ML) is not always feasible (Dryden et al., 2002). Therefore, this paper compares the inference quality using the pseudolikelihood (PML, Besag, 1974, 1977; Okabayashi et al., 2011; Sutton and McCallum, 2007), a well-known computational shortcut, to the full ML. It is generally assumed that maximum pseudolikelihood estimators (MPLEs) are defensible as reasonable alternatives to superior maximum likelihood estimators (MLEs) whenever the MLE is too difficult to compute (Besag, 1974; Sutton and McCallum, 2007; Liang and Jordan, 2008). The second purpose of this paper is to assess how the MPLE performs compared to the MLE when there is model violation in the form of non-Gaussian behavior of x. Therefore, the generalized lambda distribution (Joiner and Rosenblatt, 1971) will be simulated as well to investigate robustness to departure from the Gaussian distribution.

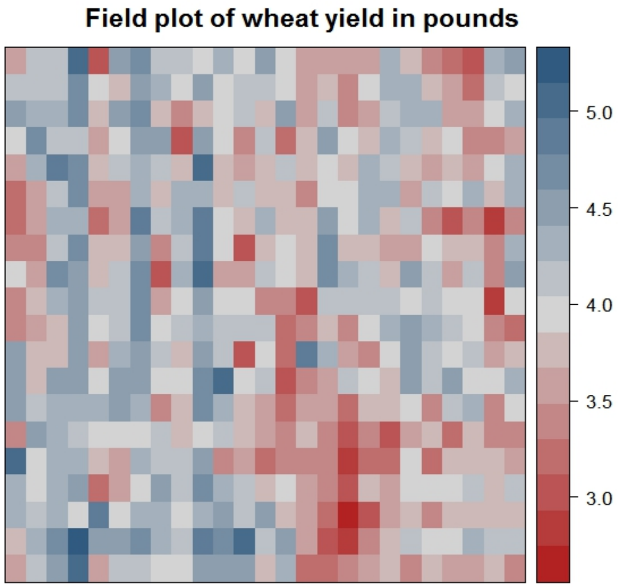

FIG 1. WHEAT YIELDS IN POUNDS ACROSS A 1 ACRE UNIFORM PLOT.

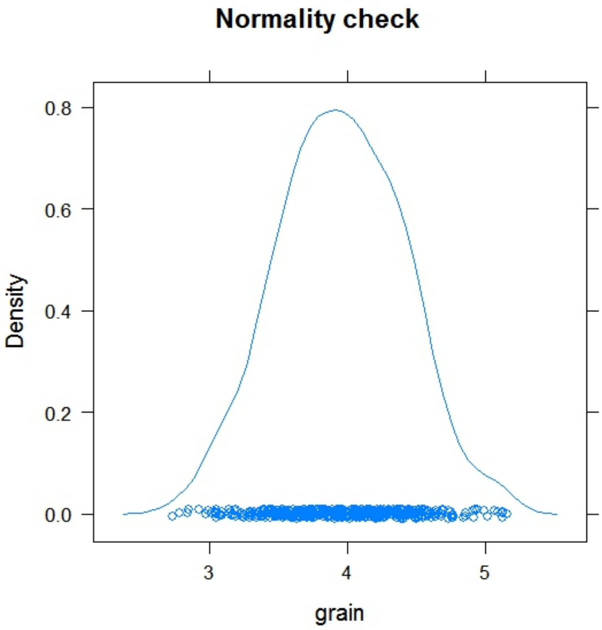

FIG 2. ESTIMATED MARGINAL PROBABILITY DENSITY ACROSS THE 500 SPATIAL INDICES (20 ROWS, 25 COLUMNS RESULTS IN 500 INDICES) IN FIGURE 1.

Data Model

A Gaussian Random Field is a finite dimensional random vector x having a multivariate Gaussian distribution, denoted $x \sim N(\mu, \Sigma)$. The vector x is a Gaussian Markov Random Field (GMRF) if it also satisfies the Markov property of conditional independence. A common approach to specify the joint density is to specify each of the conditional densities $x_i \mid x_{N_i} \sim N(\sum_{j \in N_i} \beta_{ij} x_j, \tau^2)$, where Ni is all those elements j that are neighbors of *i*. The neighbourhood Ni is defined most simply by using the precision matrix $P = \Sigma^{-1}$.

For all of our numerical examples, a size-4 neighborhood will be assumed consisiting of the east {i1+1, *i*2}, south {i1, *i*2-1}, west {i1-1, *i*2}, and north neighbor {i1, *i*2+1} of *i* = {i1, *i*2}. To avoid edge effects and simplify the calculations, we follow the convention of wrapping the image on a torus, and interpreting addition and subtraction as being modulo *n*, where *n* is the total number of rows and columns (Rue and Held, 2005). In this paper, it is assumed that $\beta_{ij} = \beta$ for all i and j. Then the precision matrix *P* has diagonal entries τ^2 and off-diagonal entries 0 in row *i* except for the columns corresponding to the neighbors of *i*, Ni, which have entries $-\beta\tau^2$. It is then a well known result regarding multivariate Gaussian distributions that x_i and x_k are conditionally independent given the four values x_j in the neighbourhood of *i*, Ni.

One challenge with moderate or large dimensional vectors x is to estimate the parameters β and τ^2. Maximum likelihood estimation (MLE) is conceptually straightforward, but computationally demanding or sometimes impossible. Besag (1974, 1977a,) showed that an approximate estimation technique based on maximizing the pseudolikelihood (MPLE) can provide an estimate, albeit a lower quality estimate (higher uncertainty) that MLEs.

A few studies have compared MLEs to MPLEs (e.g., Dryden et al., 2002) for various GMRFs, but more studies are needed using different assumptions and sample sizes in order to widen the experience base. The purpose of this paper is to compare the MLE to the MPLE for a simple 4-neighbor GMRF on a square lattice of indices denoted L_n^2. In this case, the MLE has a simple form, so the MLE is relatively fast to compute for modest-sized vectors x. The MPLE has a simple form whose terms provide insight into the terms in the precision matrix *P*. The next section gives the ML and PML estimators.

MLE and PMLE for the 4-neighbor GMRF

The MLE equations are the simplest to express if we write the probability density for *x* in the lattice of indices L_n^2 as

$$f(x)=\left|P\right|^{1/2}(2\pi)^{-n^2/2}\exp\{-\frac{1}{2}\sum_{i\in L_n^2}\sum_{j\in N_i}\theta_1(x_i-x_j)^2-\frac{\theta_2}{2}\sum_{i\in L_n^2}(x_i-\mu)^2\}\quad(1).$$

From Eq. (1) it can be shown that the MLE equations for θ_1 and θ_2 satisfy

$$\frac{1}{n^2}\sum_{k\in L_n^2}\frac{4\varpi_k}{\hat{\theta}_2+\hat{\theta}_1\varpi_k}=\frac{1}{n^2}\sum_{i\in L_n^2}\sum_{j\in N_i}(x_i-x_j)^2 \text{ and}$$

$$\frac{1}{n^2}\sum_{k\in L_n^2}\frac{1}{\hat{\theta}_2+\hat{\theta}_1\varpi_k}=\frac{1}{n^2}\sum_{i\in L_n^2}\sum_{j\in N_i}(x_i-\overline{x})^2\quad(2),$$

where $\overline{x}$ is the sample mean of the x_i and $\varpi_k=1-\cos(\frac{2\pi k_1}{n})+1-\cos(\frac{2\pi k_2}{n})$ (Burr and Kurien, 1991). The parameters θ_1 and θ_2 are hatted to denote that the solutions to Eq. (2) are estimators, the MLEs.

The pseudo likelihood is the product of the conditional distribution $x_i\,|\,x_{N_i}\sim N(\sum_{j\in N_i}\beta_{ij}x_j,\tau^2)$ (Besag, 1974). The closed-form PMLE equations for the 4-neighbor GMRF are therefore (Dryden et al., 2002)

$$\hat{\beta}_{PMLE}=\frac{\sum_{i\in L_n^2}\sum_{j\in L_n^2}(w_{i,j}x)x_{i,j}}{\sum_{i\in L_n^2}\sum_{j\in L_n^2}(w_{i,j}x)(w_{i,j}x)},$$

where $(w_{ij}x)=(x_{i,j-1}+x_{i,j+1}+x_{i-1,j}+x_{i+1,j})$ is a vector that facilitates a regression of $x_{i,j}$ on its neighbors and

$$\hat{\tau}^2_{MPLE}=\frac{1}{n^2}\sum_{i\in L_n^2}\sum_{j\in L_n^2}(x_{ij}-(w_{i,j}x)^T\hat{\beta}_{MPLE})^2\qquad(3).$$

The MLE equations must be solved iteratively, while the MPLE equations have a simple closed form and are much faster to solve. The parameters satisfy

$$\tau^2=\frac{1}{\theta_2+2\theta_1}\text{ and }\beta=\frac{\tau^2\theta_1}{2}.$$

Figure 3 is the same as Figure 1, except that Figure 3 plots a realization from the GMRF in Eq. (1) using an n-by-n square grid L_n^2 with $n^2=100$ and with $\theta_1=0.6$ and $\theta_2=1$. We use $\theta_1=0.6$ and $\theta_2=1$ in all our simulation results given in the next section.

Figure 4 plots the estimated density of the 100 x_i values, and for a sample of 100 observations, it looks reasonably close to Gaussian as it should.

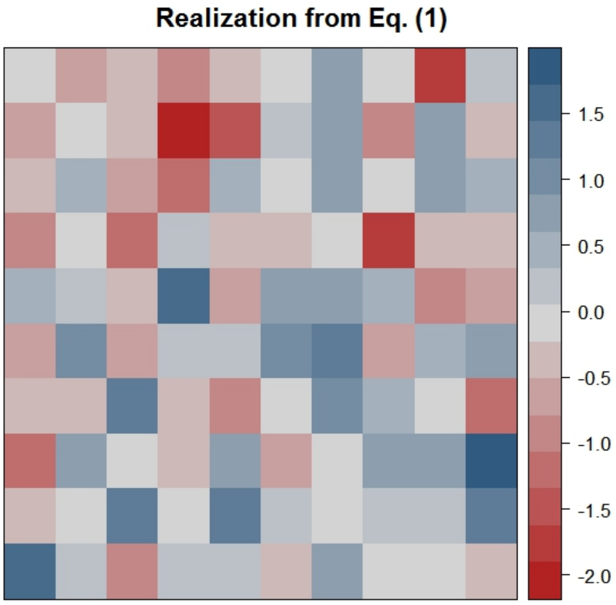

FIGURE 3. A REALIZATION FROM THE PROBABILITY DENSITY ACROSS THE 100 SPATIAL INDICES SIMULATED FROM EQ. (1).

Figure 5 is the same as Figure 3, but the marginal distribution of each xi is a generalized lambda with skewness -0.5 and kurtosis 0.7. The Gaussian skewness and kurtosis are both zero.

Figure 6 plots the estimated density of the 100 x_i values and illustrates non Gaussian behaviour.

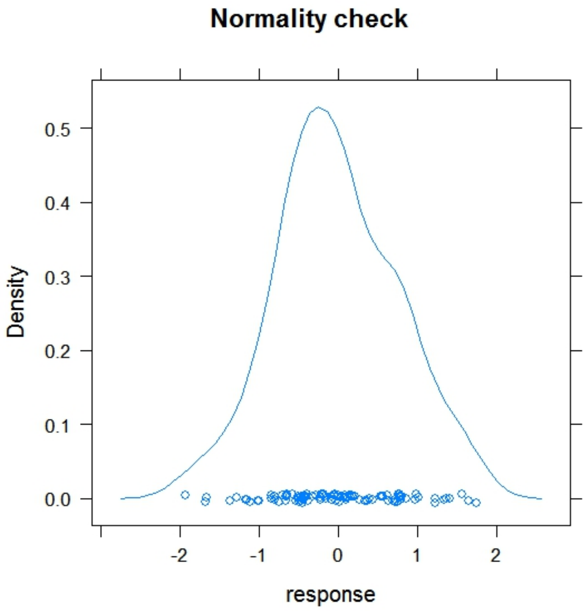

FIG 4. ESTIMATED MARGINAL PROBABILITY DENSITY FROM A REALIZATION OF DATA FROMTHE 100 SPATIAL INDICES IN FIGURE 3.

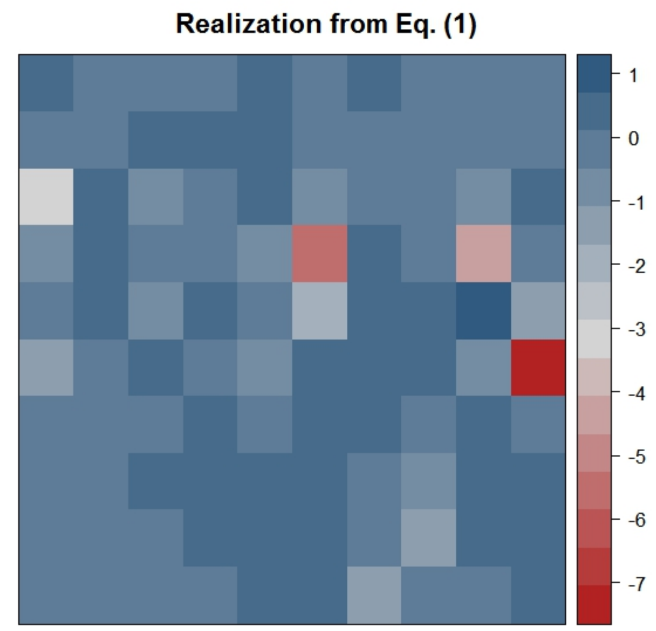

FIG 5. ESTIMATED MARGINAL PROBABILITY DENSITY ACROSS THE 100 SPATIAL INDICES SIMULATED FROM THE GENERALIZED LAMBDA DISTRIBUTION WITH COVARIANCE GIVEN BY $\Sigma = P^{-1}$ AS IN EQ. (1).

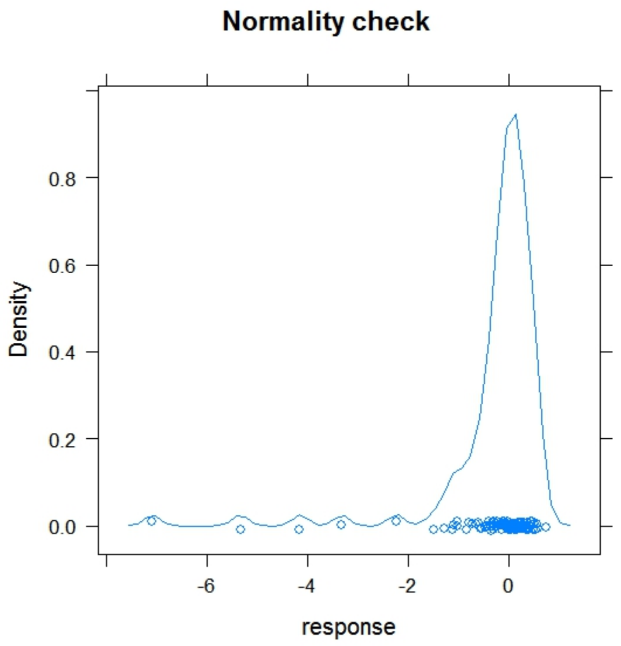

FIG 6. ESTIMATED MARGINAL PROBABILITY DENSITY ACROSS THE 100 SPATIAL INDICES IN FIGURE 5.

The next section provides simulation results comparing the mean squared error (MSE) in the MPL estimator to the ML estimator for a range of vector sizes, first assuming that the normal distribution is exactly correct and then assuming there is mild departure from normality as captured in the generalized lambda distribution which can have non-Gaussian skewness and/or kurtosis.

The MLE is known to be asymptotically efficient, which means that it has the smallest variance as the lattice size increases. Of course, for large sample size, any good estimation method should have reasonably small variance. In practice, as the lattice size increases, alternatives to the simple 4-neighbor GMRF could be considered.

Simulation Results

All of our analyses are for simulated data from the probability density in Eq. (1) or for simulated data that have the same covariance as in Eq. (1), but have non-Gaussian marginal distributions. The wheat data was presented in Figure 1 and 2 for context and background, but it has been analysed elsewhere (Gelfand et al., 2010). Model selection is an important task in analysis of any real data set (Kaiser et al., 2012). Our focus is on comparing the performance of MPLE to that of MLE assuming that we know the correct form of covariance matrix, leaving only the task of parameter estimation, not the broader task of model selection.

We implemented the MLE and PMLE equations in R (R, 2004) and simulated x on a grid of values n = 5, 7, 10, 15, 20, and 30. For ease of comparison, we estimated τ^2 and β for PMLE and for MLE. The MLEs for τ^2 and β are computed by first using optim in R (with the Nelder and Mead algorithm) to solve the MLEs in Eq. (2) and then using $\hat{\tau}^2 = \dfrac{1}{\hat{\theta}_2 + 2\hat{\theta}_1}$ and $\hat{\beta} = \dfrac{\hat{\tau}^2\hat{\theta}_1}{2}$.

Table 1 gives the estimated root mean squared error (RMSE) in 100 simulations (repeatable to within approximately ±10% of the table entry, which is verified by repeating twice each set of $nsim$ = 100 simulations) for the MLE and the PMLE for each value of n. The RMSE is calculated for example, for $\hat{\beta}$, as

$$RMSE = \sqrt{\frac{\sum_{i=1}^{nsim} (\hat{\beta} - \beta)^2}{nsim}}.$$

Table 2 is the same as Table 1, but the marginal distribution of each x_i was a generalized lambda distribution (gldist in R using skewness parameter 0.5 and kurtosis parameter 0.7 as in Figures 5 and 6) and the Cholesky factorization of Σ was used to convert the joint distribution of the x_i to have the desired precision matrix *P*.

Table 3 is the same as Table 2, but the skewness and kurtosis parameters were -0.5 and 0.7. The generalized lambda distribution (Joiner and Rosenblatt, 1971) is a convenient distribution that includes the Gaussian as one special case, and that allows for non-Gaussian skewness and/or kurtosis (tail behaviour).

TABLE 1 RMSES FOR GAUSSIAN DATA FOR $(\hat{\beta}, \hat{\tau}^2)$. ENTRIES OF 0.01 OR LESS ARE RECORDED AS 0.01

Lattice Size *n*	MLE	MPLE
5	0.25, 0.05	0.13, 0.20
7	0.24, 0.05	0.10, 0.14
10	0.18, 0.03	0.07, 0.08
15	0.09, 0.02	0.04, 0.06
20	0.05, 0.01	0.03, 0.04
30	0.03, 0.01	0.02, 0.03

TABLE 2 RMSES FOR NON-GAUSSIAN DATA

Lattice Size *n*	MLE	MPLE
5	0.29, 0.06	0.14, 0.19
7	0.26, 0.05	0.26, 0.13
10	0.23, 0.05	0.22, 0.09
15	0.12, 0.03	0.10, 0.05
20	0.11, 0.03	0.11,0.04
30	0.08, 0.02	0.08, 0.03

TABLE 3 RMSES FOR NON-GAUSSIAN DATA

Lattice Size *n*	MLE	MPLE
5	0.32, 0.07	0.46, 0.18
7	0.28, 0.06	0.29, 0.12
10	0.23, 0.05	0.23, 0.09
15	0.15, 0.03	0.15, 0.05
20	0.10,0.02	0.10,0.04
30	0.07, 0.02	0.07, 0.03

Notice in Table 1 that the MLE has smaller RMSE than the MPLE for $\hat{\tau}^2$ and larger RMSE than the MPLE for $\hat{\beta}$ until the grid size is large, approximately 30 or more. It was verified that the larger RMSE in the MLE for small sample sizes is due to bias. The MLE is asymptotically unbiased (Burr and Kurien, 1991), but apparently the bias is nonnegligible for small sample sizes.

Notice in Table 2 that for non-normal data having a generalized lambda (skewness 0.5 and kursotis 0.7), the RMSEs generally increase more the MPLE than for MLE. Interestingly, the smaller RMSEs for MPLE in Table 1 for $\hat{\beta}$ are substantially increased in Table 2, suggesting that the advantage of MPLE for small samples requires close to Gaussian data.The trends in RMSEs in Table 3 for non-normal data having a generalized lambda (skewness -0.5 and kursotis 0.7) are very similar to the trends in Table 2.

Conclusions and Summary

When the Gaussian distribution holds exactly, the MPLE has surprisingly small RMSE compared to the RMSE of the MLE for $\hat{\beta}$ and reasonably small RMSE for $\hat{\tau}^2$.

GMRFs are often applied to superpixels which are patches of similar pixels. Therefore, images consisting of hundreds of thousands of raw pixels are often reduced in preprocessing to a more modest number of pixels. Our simulation experiment found that for lattice sizes of 30 or more, the RMSE of the MPLE is essentially the same as the RMSE of the MLE.

When the marginal distributions are generalized lambda with skewness 0.5 or -0.5 and kursotis 0.7, the RMSEs increase more for MPLE than that for MLE. Model selection of a more appropriate likelihood is therefore recommended in future work in which we will experiment with the composite likelihood (Varin et al., 2011) that includes pseudolikelihood or piecewise pseudoliklihood (Sutton and McCallum, 2007) as special cases and is another alternative to ML.

ACKNOWLEDGMENT

We acknowledge the support of the U.S. Department of Energy through the LANL/LDRD Program for this work.

REFERENCES

Besag, J., "Spatial Interaction and the Statistical Analysis of Lattice Systems," Journal of the Royal Statistical Society B, 36(2), pp. 192-225, 1974.

Besag, J., "Efficiency of Pseudolikelihood Estimation for Simple Gaussian Fields," Biometrika 64, pp. 616-618, 1977.

Burr, T., Kurien, T., "Estimating and Modeling Gene Flow for a Spatially Distributed Species, " Florida State University Technical Report M 837, 1991.

Dryden, I., Ippolitii, L., Romagnoli, L., "Adjusted Maximum Likelihood and Pseudo-likelihod Estimation for Noisy Gaussian Markov Random Fields," Journal of Computational and Graphical Statistics, 11:2, pp. 370-388, 2002.

Gelfand, A., Diggle, P,. Fuentes, M., Guttorp, P, "Handbook of Spatial Statistics," Chapman and Hall, Boca Raton, 2010.

Hastie, T., Tibshirani, R., Friedman, J., "The Elements of Statistical Learning," Springer, New York, 2001.

Kaiser, M., Soumendra, L., Nordman, D., “Goodness of Fit Tests for a Class of Markov Random Field Models,” Annals of Statistics 40(1), pp. 104-130, 2012.

Mardia, K., Kent, J., Hughes, G., Taylor, C., “Maximum Likelihood Estimation Using Composite Likelihoods for Closed Exponential Families,” Biometrika 96, pp. 975-982, 2009.

Joiner, B., Rosenblatt, J., “Some Properties of the Range in Samples from Tukey's Symmetric Lambda Distributions,” Journal of the American Statistical Association, 66(334), pp. 394-399, 1971.

Liang, P, Jordan, M. “An Asymptotic Analysis of Generative, Discriminative, and Pseudolilkelihood Estimators,” Proceedings 25th International Conference on Machine Learning, 2008.

Okabayashi, S., Johnson, L, Geyer, C., “Extending Pseudo-likelihood for Potts Models,” Statistica Sinica 21, pp. 331-347, 2011.

R Development Core Team. R: a Language and Environment for Statistical Computing. R Foundation for Statistical Computing, www.r-project.org, 2004.

Rue, H. Held, L., “Gaussian Markov Random Fields Theory and Applications,” Chapman and Hall, Boca Raton, 2005.

Sutton, C. McCallum, A., “Piecewise Pseudolikelihood for Efficient Training of Conditional Random Fields,” Computer Sciences Department Faculty Publication Series, paper 62, 2007.

Varin, C., Reed, N., Firth, D., “An Overview of Composite Likelihood Methods,” Statistica Sinica 21, pp. 5-42, 2011.